ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Decision Support Systems 37 (2004) 175–186

Decision Support Systems

www.elsevier.com/locate/dsw

Feedback-labelling synergies in judgmental stock price forecasting

Paul Goodwin[a,*], Dilek Önkal-Atay[b,1], Mary E. Thomson[c,2], Andrew C. Pollock[d,3], Alex Macaulay[d,4]

[a] *The Management School, University of Bath, Claverton Down, Bath BA2 7AY, UK*
[b] *Faculty of Business Administration, Bilkent University, Ankara 06533, Turkey*
[c] *Glasgow Caledonian Business School, Cowcaddens Road, Glasgow G4 0BA, UK*
[d] *Division of Mathematics, School of Computing and Mathematical Sciences, Glasgow Caledonian University, Cowcaddens Road, Glasgow G4 0BA, UK*

Received 1 October 2002; accepted 5 December 2002

Abstract

Research has suggested that outcome feedback is less effective than other forms of feedback in promoting learning by users of decision support systems. However, if circumstances can be identified where the effectiveness of outcome feedback can be improved, this offers considerable advantages, given its lower computational demands, ease of understanding and immediacy. An experiment in stock price forecasting was used to compare the effectiveness of outcome and performance feedback: (i) when different forms of probability forecast were required, and (ii) with and without the presence of contextual information provided as labels. For interval forecasts, the effectiveness of outcome feedback came close to that of performance feedback, as long as labels were provided. For directional probability forecasts, outcome feedback was not effective, even if labels were supplied. Implications are discussed and future research directions are suggested.

Keywords: Forecasting; Judgment; Feedback; Calibration; Stock price; Contextual information

1. Introduction

Forecasting and decision support systems are partly systems for learning. One of their objectives is to improve management judgment by fostering understanding and insights and by allowing appropriate access to relevant information [16]. Feedback is the key information element of systems that are intended to help users to learn. By providing managers with timely feedback, it is hoped that they will learn about the deficiencies in their current judgmental strategies and hence enhance these strategies over time. When a system is being used to support forecasting, feedback can be provided in a number of forms [6,10]. The simplest form is *outcome feedback*, where the manager is simply informed of the actual outcome of an event that was being forecasted. *Performance feedback* provides the forecaster with a measure of his or her

* Corresponding author. Tel.: +44-122-538-3594; fax: +44-122-582-6473.
E-mail addresses: mnspg@management.bath.ac.uk (P. Goodwin), onkal@bilkent.edu.tr (D. Önkal-Atay), mwi@gcuexch.gcal.ac.uk (M.E. Thomson), a.c.pollock@gcal.ac.uk (A.C. Pollock), A.B.Macaulay@gcal.ac.uk (A. Macaulay).
[1] Tel.: +90-312-290-1596; fax: +90-312-266-4958.
[2] Tel.: +44-141-331-8954; fax: +44-141-331-3229.
[3] Tel.: +44-141-331-3613; fax: +44-141-331-3608.
[4] Tel.: +44-141-331-3052; fax: +44-141-331-3608.

0167-9236/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-9236(03)00002-2

forecasting accuracy or bias. *Process feedback* involves the estimation of a model of the forecaster's judgmental strategy. By feeding this model back to the forecaster, it is hoped that insights will be gained into possible ways of improving this strategy. Finally, *task properties feedback* delivers statistical information on the forecasting task (e.g. it may provide statistical measures of trends or correlations between the forecast variable and independent variables).

Most of the research literature on management judgment under uncertainty suggests that outcome feedback is less effective than other forms in promoting learning (e.g. Refs. [6,33]). For example, much research into the accuracy of judgmental forecasts has found that forecasters tend to focus too much on the latest observation (e.g. the latest stock value) which will inevitably contain noise. The result is that they see evidence of new, but false, systematic patterns in the latest observation [31] and overreact to it. Because outcome feedback draws attention to the latest observation it exacerbates this tendency. This means that a long series of trials may be needed to distinguish between the systematic and random elements of the information received by the forecaster [31].[5] In contrast, by averaging results over more than one period (or over more than one series if cross-sectional data is being used), other forms of feedback are likely to reduce the attention that is paid to the most recent observation and to filter out the noise from the feedback. For example, performance feedback may be presented in the form of the mean forecast error, or in the case of categorical forecasts, the percentage of forecasts that were correct.

However, if conditions could be found where outcome feedback does encourage learning as efficiently (or nearly as efficiently) as other forms of feedback, then this would yield considerable benefits to users and designers of support systems. This is because outcome feedback overcomes, or at least reduces, various shortcomings of the other forms.

Firstly, outcome feedback is easier to provide and is likely to be more easily understood by the forecaster. Conversely, the provision of performance feedback, for instance, can involve difficult choices on which performance measure to provide—each measure will only relate to one aspect of performance, but providing several measures may confuse the forecaster. Moreover, some measures may be difficult to comprehend and will therefore require that the forecaster is trained in their use. Process feedback will require the identification of cues that the forecaster is assumed to be using, with no guarantee that these cues have really been used. Also, multicollinearity in these cues means that there will be large standard errors associated with the estimates of the weights that the forecaster is attaching to the cues. Task properties feedback requires regular statistical patterns in past data. By definition, these characteristics are often absent in tasks where management judgment is preferred to statistical methods.

Secondly, when judgments are being made in relation to a *single variable over time*, outcome feedback will not be contaminated by old observations when circumstances are changing. Because performance and process feedback are measured over a number of periods, they may lag behind changing performance or changes in the strategies being used by the forecaster. Also, several periods must elapse before a meaningful measure of performance, or a reliable model of the judgmental process, can be obtained. For *cross-sectional data*, outcome feedback can be provided for each variable and, as such, is not merely an average of potentially different performances (or strategies) on different types of series. Furthermore, a reasonably large number of judgments over different series are required in order to obtain reliable estimates of performance or a reliable estimate of the process model.

As we discuss below, there are some indications in the literature of situations that *may* be favourable to outcome feedback. These relate to (i) the nature of the forecast that is required, and (ii) the type of information that is supplied with the feedback—in particular, whether the past history of the forecast variable is accompanied by an informative label.

This paper describes an experiment that was used to investigate the effects of these factors in an important application area: stock price forecasting. Financial forecasting is an area where human judgment is particularly prevalent [8,35,45] and the specific role

[5] It is possible that, in some circumstances, outcome feedback is actually damaging to the quality of judgments. However, since in most practical forecasting tasks it will be difficult to avoid the forecaster having access to outcome feedback, the identification of factors that will mitigate its effects would then be of interest.

of judgment in forecasting stock prices has itself received particular attention from the research community (see Refs. [7,21,26,32–34,41,48]). The paper compares the effectiveness of outcome feedback under different conditions with that of performance feedback. Performance feedback was used as the benchmark because, of the other feedback types, it is likely to be the most relevant to financial forecasting and most acceptable to forecasters. The paper is structured as follows. First, a literature review is used to explain why outcome feedback may be more effective when particular types of forecasts are required and why feedback type and label provision might be expected to have interactive effects. Then details of the experiment are discussed, followed by analysis and discussion. The paper concludes with suggestions for further research.

2. Literature review

2.1. Feedback and type of forecast

There is some evidence in the literature that the effectiveness of outcome feedback is related to the nature of the forecast that is required. It seems that outcome feedback is unlikely to be effective when point forecasts are required (e.g. Ref. [24]). *Point forecasts* merely provide an estimate of the exact value that the forecast variable will assume at a specified time in the future (e.g. the stock price of company X at the end of trading tomorrow will be $3). As indicated earlier, this is probably because outcome feedback exacerbates the tendency to read system into the noise that is associated with the most recent observation. Point forecasts fail to communicate the level of uncertainty that is associated with the forecast. In contrast, judgmental *interval forecasts* (e.g. "I am 90% confident that the closing stock price will be between $2.4 and $3.6") do indicate the level of uncertainty, and there is some evidence that outcome feedback is effective in improving these. Usually, the estimated intervals are too narrow for the specified level of confidence, but a study by O'Connor and Lawrence [29] found that outcome feedback was effective in widening the intervals. This may be because the difference between the reported outcome and the original forecast draws attention to the inherent uncertainty associated with the forecasting task. There is also some evidence that *categorical probability forecasts* (e.g. the probability that it will rain during the next 24 h) can be improved by outcome feedback. Indeed, the almost perfect calibration of US weather forecasters has been partly attributed to the fact that the forecasters receive regular and speedy outcome feedback relating to their forecasts [3]. *Directional probability forecasts* (e.g. "I am 80% confident that the stock price at the end of trading in seven days time will be lower than the current price")[6] can be regarded as a special case of categorical probability forecasts and may therefore also benefit from outcome feedback. However, in the financial forecasting context, Önkal and Muradoglu [33] have shown performance feedback to be more effective than simple outcome feedback in improving the accuracy of stock price forecasts expressed as probabilities over multiple price-change intervals.

2.2. The effect of providing labels

In helping the forecaster, a support system can provide various levels and types of information. *Time series information* indicates the past history of the forecast variable, enabling trends or other patterns to be identified and the volatility of the variable to be assessed. *Contextual information* refers to information about the forecast variable over and above the series history. For example, it might refer to information of a company takeover. It also includes *labels*, which simply indicate the nature of the series (e.g. the name of the company whose past stock prices are being displayed). As we indicate below, research suggests that labels can have a profound effect on judgmental forecasts. It is also notable that many financial forecasters base their estimates *only* on time series information (i.e. the use of specific labels is absent). For instance, chartists do not use any contextual information due to their belief that all indicators of change (i.e. economic, political, psychological or otherwise) are reflected in the pattern of the price series itself and, therefore, a study of past price movements is all that is needed to forecast future movements [27,28].

Labels are a particularly interesting form of contextual information that can have powerful effects on

[6] This type of forecast is preferred over the multiple-interval format by both financial professionals [44] and theorists [17].

the accuracy of judgmental forecasts. These effects can occur because labels create expectations about the form and nature of the time series [39]. Labels which create expectations that are congruent with the statistical structure of the task can improve the accuracy of prediction because they increase knowledge of this structure. They also improve the consistency of prediction because they reduce the need to search for and test a large variety of hypotheses about the nature of the data. However, Sniezek [39] found that even neutral labels (i.e. labels which give a context to the task, but give no information about statistical structure, such as "weather" and "marketing variable") can aid performance, possibly because they offer a context to a task that allows the judge to create, with the observed data, a congruent, and hence consistent, interpretation. Even non-expert forecasters may benefit from labels in this way.

A concept that is closely related to congruence is label specificity. For example a graph can simply display the general label "sales" or it could display the more specific label "sales of mobile phones by the Acme phone company". It is possible that specificity can have a profound effect on the way a series is interpreted. Beach et al [9] have distinguished between the use of aleatory and epistemic strategies in judgmental forecasting. Aleatory strategies categorise elements by their class membership, rather than their unique properties (e.g. given that you are a member of a particular profession, you have a 70% probability of living to an age of 80 or more). If only a general label like "sales" is presented, a time series can only be seen as a member of the class of sales time series which may be perceived to behave in a stereotypical way. For example, consider the use of the label "sales" in a judgmental forecasting study by Lawrence and Makridakis [19]. Despite the fact that graphs of the time series manifested an upward linear trend, the label may have caused subjects to forecast damped growth, because sales series typically have this pattern. Hence the use of general labels may cause forecasters to pay less attention to the specific characteristics of the series and more to the perceived characteristics of the stereotypical pattern. In forecasting stock prices, this may involve perceptions such as "recent gains are usually subsequently reduced by profit taking". Nonspecialist forecasters may not have such perceptions and they may have difficulty in making any sense of movements in the stock price time series.

In contrast, epistemic strategies use information on the unique characteristics of the element in question (e.g. you are 30 years old, eat healthily, exercise regularly, do not smoke etc., so you have 85% chance of living to an age of 80 or more). Providing a specific label might therefore be expected to promote epistemic reasoning with a greater focus on the individual features of the time series. This may be beneficial if it enables the forecaster to incorporate company-specific knowledge into the interpretation of the graph and the forecast—with 'important' movements in the time series being more salient and more meaningful. For example, in stock price forecasting, this may involve considerations like "this company is in the aerospace industry and given recent bad news about this industry I expect the slight downward movement in the share price to continue". It will be detrimental if it encourages the forecaster to attempt to explain specific movements in the series that are best regarded as noise [13].

Finally, Beach et al. [9] have also suggested that one of the determinants of the motivation to produce accurate forecasts is the quality and amount of information available to the forecaster—other things being equal, the more adequate the information, the greater the expectation of forecasting accuracy. Hence, the motivation for accurate forecasts may be expected to increase if 'general labels' are replaced by more specific labels. Again, this means that even non-expert subjects may be expected to improve their performance as the specificity of the labels increases.

2.3. The interaction between labels and feedback

One important area that has been underexplored in the literature is the possible *interaction* between feedback types and the extent of availability of contextual information. Yet such interactions may be of considerable interest. For example, in a cue probability learning task, Adelman [1] found that the provision of congruent labels led to no difference in performance in a cue probability learning task between task properties and outcome feedback. Adelman suggested that this may have resulted because the labels implicitly provided accurate information about the statistical structure of the task thereby matching the information

that was explicitly provided by the task properties feedback. It is thus possible that, by providing advance information about the data, meaningful labels can add to task knowledge and enhance the rate of learning that would be achieved by feedback alone.

In a judgmental time series forecasting task, it is possible to hypothesise about the effect of interactions between the specificity of the label provided and the type of feedback. Both the labels and the feedback can be viewed in terms of their likely effects on the attention that the judgmental forecaster will pay to the time series pattern. At one extreme, the absence of specific labels and the provision of only outcome feedback are both likely to reduce the salience of the overall time series pattern provided and encourage a focus on the most recent value. When specific labels are provided *with* the outcome feedback, they are likely to increase the salience and meaningfulness of the particular time series pattern and hence improve forecast accuracy. In particular, considering the entire time series should improve the forecaster's assessment of the amount of uncertainty associated with the forecast variable and hence improve interval forecasts so that their width is more appropriate for the level of confidence that is being expressed.

Where performance feedback is a summary measure taken across a number of time series, it should serve to alert the forecaster to general deficiencies in his or her forecasting strategy and engender reflection on how improvements might be achieved. This would also encourage the forecaster to attend to the entire time series pattern, even when no specific labels are supplied (and even where outcome feedback is also provided) and may account for the benefits of performance feedback reported in the Önkal and Muradoglu study [33]. The interesting question is whether providing specific labels yields any added value in the presence of performance feedback. If both specific labels and performance feedback improve performance by directing attention to the overall time series pattern, then the effect of one of these information types may be subsumed within that of the other.

3. Method

Participants were undergraduate business students from Bilkent University who were taking a forecasting course. Participation was voluntary and no compensation was provided. The subjects were randomly assigned to four groups based on the type of feedback (outcome vs. performance feedback) and the provision of labels (names of stocks provided vs. not provided). Performance feedback was provided in the form of calibration feedback. Calibration refers to the correspondence between the forecaster's probabilities and the attained proportion of correct forecasts. For example, if a perfectly calibrated forecaster is expressing his or her forecasts as 90% prediction intervals, we would expect 90% of these intervals to include the true value of the variable being forecast. Similarly, when this forecaster states that her/she is 80% confident that a stock price will move in a particular direction, we would expect the predicted direction to be correct on 80% of occasions. Calibration is an integral aspect of performance and is therefore a natural choice for performance feedback (for extensive reviews of this literature, see Refs. [22,25]). Note that *all* subjects were provided with outcome feedback in order to provide a realistic simulation of stock market forecasting. In a practical situation, it is very unlikely that a forecast would be made without the forecaster having knowledge of the most recent observation. Moreover, the denial of access to this information would mean that the task became progressively more difficult as forecasters were forced to make forecasts with increasing lead times. It would therefore mean that lead time was confounded with absence of outcome feedback in the experimental design. Note also that the designation "no-labels" (below) means that "no specific labels" were provided since a general label "stock prices" is implied by the nature of the task.

Fifty-nine students completed the 3-week long experiment. In particular, the groups were organized as follows:

G1: outcome feedback, no-labels group ($n = 14$),
G2: outcome feedback, labels group ($n = 12$),
G3: calibration feedback, no-labels group ($n = 17$),
G4: calibration feedback, labels group ($n = 16$).

For each of the three sessions, participants were requested to provide weekly interval and probability forecasts for the closing stock prices of 30 randomly selected companies from the Istanbul Stock Exchange.

- I am 90% confident that the closing value of this stock price that will be realized on Friday will be between _____ and _____.

- When compared to the previous Friday's closing price, the coming Friday's closing price will

  A. Increase
  B. Stay the same or decrease

Your forecast (A or B) : _____

Probability that your forecast will indeed occur
(i.e., probability that the weekly change will actually fall
in the direction you predicted) (BETWEEN 50% AND 100%) : _____

Fig. 1. Sample form for reporting judgmental forecasts.

Selection of the weekly forecast horizon was dictated by the conditions prevailing in emerging markets. Ordering of 30 stocks was randomized individually for each subject for each session. All subjects were given the weekly closing stock prices (i.e. the closing stock prices for each Friday) for the previous 52 weeks in graphical form; and, so that subjects had appropriate information to provide numerical values for credible intervals, the data were also presented for the previous 12 weeks in tabular form. The name of each stock was provided to the subjects in the labels groups (i.e. G2 and G4), whereas the stock names were not revealed to the subjects in the no-labels groups (i.e. G1 and G3). At the beginning of second and third sessions, participants in G1 and G2 received outcome feedback (i.e. previous Friday's closing prices marked on the graphical and tabular information forms). Subjects in G3 and G4 received calibration feedback in addition to the outcome feedback. Specifically, subjects in groups 3 and 4 were given (1) closing prices of the previous week shown on the tabular and graphical forms; (2) individual calibration scores computed from the previous week's probability forecasts, along with detailed information on the proportion of correct forecasts and relative frequency of use for each probability category employed by the participant; and (3) percentage of their prediction intervals that actually contained the realized stock price (i.e. an index of interval calibration).

At the beginning of the first session, concepts of "subjective probability", "prediction intervals" and "probability forecasting" were discussed, and their role in financial forecasting was emphasized. Examples were given and the participants were informed that certain scores of forecasting performance would be computed from their individual forecasts, and that they could earn their best potential score by stating their true opinions without hedging or bluffing. Also, the students in no-contextual-information groups were specifically instructed to base their forecasts only on the price information presented, without trying to uncover the names of individual stocks. These subjects were warned of the particular significance of basing their forecasts solely on the presented time series information.

In each session, the participants were instructed to provide a prediction interval for the closing price of each of the stocks being considered. In stating the prediction interval, each subject gave the highest and the lowest predicted closing price for each stock such that he/she was 90% confident that this range would include the actual closing price. Participants were also asked to make directional probability forecasts for the closing prices of stocks. In particular, each subject was requested to indicate whether (s)he believed that the stock price for the current Friday would (a) increase, or (b) decrease or stay the same in comparison with the previous Friday's closing stock price. Following this direction indication, each subject was asked to convey his/her degree of belief with a subjective probability for the predicted direction of price change (i.e., probability that the weekly price change would actually fall in the direction indicated by the subject). Since a direction of price change was given first, the following probability would have to lie between 0.5 and 1.0. A sample form for reporting predictions is presented in Fig. 1.

4. Findings

The participants' forecasts were evaluated using two performance measures: hit rate (HR) and root mean probability score (RMPS). Details of the performance measures are provided in Appendix A.[7] Hit Rate (HR) refers to the percentage of intervals that include the actual value, and is commonly used as an index of interval calibration [2,22]. As indicated earlier, a set of 90% prediction intervals would be well-calibrated if the actual values fell within these intervals 90% of the time. Overconfidence is exhibited if less than 90% of the realized values fall within the specified intervals; underconfidence is inferred if more than 90% of the occurring values fall inside the intervals. The overall accuracy of directional probability forecasts is indexed via the root mean probability score (RMPS), with *lower* scores indicating better performance.

Table 1 provides the analysis of variance (ANOVA) results for interval and directional forecasts, from a split-plot design with two between-subjects factors [viz., *feedback* (calibration vs. outcome feedback) and *labels* (labels provided vs. not provided)], and one repeated measure within-subject factor [viz., *session* (forecasting sessions 1, 2 or 3)]. For each of the two ANOVAs, the response variable was taken as the mean, across the 30 stocks, of the corresponding performance measure (details of the assumed model and the analyses regarding the normality assumption are presented in Appendix B). The mean scores for the two the performance measures for the different factor levels are displayed in Table 2.

Table 1

	F-statistic
(a) ANOVA results for interval forecasts: hit rate	
Feedback	29.73***
Labels	13.53***
Session	72.85***
Feedback × Labels	7.34**
Feedback × Session	4.35*
Labels × Session	1.11
Feedback × Labels × Session	0.18
Normality test	0.063
(b) ANOVA results for directional probability forecasts: root mean probability score (RMPS)	
Feedback	3.25
Labels	2.02
Session	7.88***
Feedback × Labels	2.35
Feedback × Session	1.59
Labels × Session	0.07
Feedback × Labels × Session	1.80
Normality test	0.062

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

4.1. Interval forecasts

Tables 1 and 2 depict the importance of feedback type for interval calibration. That is, when participants are given calibration feedback, the attained hit rates more closely approach the specified confidence coefficient of 90%, as compared with participants given outcome feedback [$F(1,100) = 29.73$, $p < 0.001$]. In addition, higher hit rates are clearly obtained by participants given labels, as opposed to those participants not given label information [$F(1,100) = 13.53$, $p < 0.001$]. Forecasting session also shows a significant effect [$F(2,110) = 72.85$, $p < 0.001$]. However, these main effects have to be interpreted with caution, since there appear significant interactions for feedback by labels [$F(1,110) = 7.34$, $p = 0.008$] and feedback by session [$F(2,110) = 4.35$, $p = 0.015$]. In particular, as illustrated in Table 2, calibration-feedback group subjects seem to attain similar hit rates regardless of whether they are or are not provided with label information. Participants receiving only outcome feedback, on the other hand, get lower hit rates when provided with no stock names, while obtaining higher hit rates if they are provided with such labels. Also, as can be gleaned from Table 2, although both feedback groups start with relatively similar hit rates in session 1, there seem to be wider differences in pursuing sessions, with calibration-feedback group subjects securing higher hit rates in sessions 2 and 3, as compared to subjects in the outcome-feedback group. The outcome feedback, no-labels group clearly shows the lowest hit rates.

[7] Eight other performance measures were also calculated but, for brevity, these have not been reported here. These measures were as follows: mean interval profitability score, mean probability response, proportion of correctly predicted directions, root calibration, bias, slope, root scatter and mean probability profitability score. Details of these measures, and the performance of participants on them, are available from the authors.

Table 2
Mean performance scores

Forecast type	Performance measures	
	Interval hit rate |90%|	Directional RMPS ↓
Calibration feedback and labels		
Session 1	60.06%	0.545
Session 2	55.04%	0.508
Session 3	87.08%	0.513
Calibration feedback and no labels		
Session 1	53.60%	0.559
Session 2	57.84%	0.585
Session 3	84.48%	0.519
Outcome feedback and labels		
Session 1	61.27%	0.512
Session 2	43.72%	0.444
Session 3	79.44%	0.510
Outcome feedback and no labels		
Session 1	46.59%	0.517
Session 2	34.09%	0.504
Session 3	62.38%	0.534
Calibration feedback—all		
Session 1	56.83%	0.552
Session 2	56.44%	0.497
Session 3	85.78%	0.516
Outcome feedback—all		
Session 1	53.93%	0.515
Session 2	38.90%	0.574
Session 3	70.91%	0.522
Labels—all		
Session 1	60.66%	0.533
Session 2	50.38%	0.485
Session 3	83.26%	0.519
No labels—all		
Session 1	50.09%	0.529
Session 2	45.96%	0.476
Session 3	73.43%	0.511
All		
Session 1	55.38%	0.533
Session 2	47.67%	0.485
Session 3	78.35%	0.519

|90%|: Values near 90% better.
↓: Smaller values better.

4.2. Directional probability forecasts

Performance in directional probability forecasting is also affected by feedback type and session, with no evident influences of label information. That is, as shown in Tables 1 and 2, forecasting session affects the overall accuracy (i.e. RMPS) [$F(2,110)=7.88$, $p=0.001$] of judgmental probability forecasts. For the performance-feedback groups, the first session appears to have the worst performance followed by notable improvements in the second session. In the third forecasting session, the RMPS worsens (but still demonstrates better performance than that of the first session). However, over the three sessions, the performance of the outcome-feedback groups does not improve so there is no evidence that the feedback is fostering learning.

5. Conclusion and directions for future research

This research examined the effects of performance and outcome feedback on judgmental forecasting performance conditional on (i) the availability of contextual information provided in the form of labels and (ii) the form in which the forecast was expressed. Using stock prices as the forecast variables of interest, the current study employed judgmental prediction intervals and probability forecasts as formal expressions conveying the forecasters' uncertainties.

Earlier work utilizing prediction intervals in other domains has indicated that the assessors typically provide narrow intervals [15,19,20,29,30,36,40,47]. Our findings from initial experimental sessions confirm earlier results in that the participants' intervals enveloped the realized value less frequently than the desired level (i.e. 90% for the current study). In response to recurrent feedback, however, subjects were able to widen their intervals, attaining significant improvements after two feedback sessions. In particular, subjects receiving interval calibration feedback secured hit rates very close to 90% in the third session, followed by outcome-feedback groups with significantly improved, but still trailing, hit rates. This is consistent with Hammond's [14] assertion that learning through outcome feedback requires more trials than other forms of feedback as judges seek to distinguish between the systematic and random components of the outcome information.

While these findings highlight the effectiveness of interval calibration feedback on reducing interval overconfidence, they also show that simple outcome feedback is most effective when labels are provided. Indeed, by the third session, calibration in the outcome

feedback–labels condition was approaching that of the calibration-feedback conditions. As hypothesised earlier, this may have resulted from increased propensity of subjects to consider the characteristics of the entire time series pattern, rather than just the most recent value, when they were provided with company specific labels. For the calibration-feedback group, this beneficial effect may already have been achieved by providing the feedback so that the specific labels brought no added benefits to the task. This implies that in a task where only interval forecasts are required the benefits of outcome feedback that were referred to earlier (e.g. ease of provision and adaptability to new conditions) may outweigh its slightly worse performance as an aid to improving calibration, as long as specific labels are provided.

Analysis of directional probability forecasts also reflects that, even though the calibration-feedback participants displayed quite poor calibration in their forecasts of session 1, detailed feedback immediately enhanced their performance in sessions 2 and 3. In contrast, the outcome-feedback subjects maintained a relatively more uniform calibration performance throughout the sessions. These results are in agreement with Lim and O'Connor's [24] findings with point forecasts. These authors suggest that individuals may feel overconfident about their ability to acquire all the information they need from time series anyway, leading them to disregard any new negative outcome feedback. Our findings may denote that this unwarranted confidence may persist with outcome feedback, but may be overcome if detailed performance feedback is provided. In contrast to the results on interval forecasting, outcome feedback cannot therefore be recommended as an aid to learning when directional probability forecasts are required, even if labels are provided.

In fact, no significant effects of label information on directional probability forecasting performance were found. One potential explanation could be that feedback was given preeminent importance, leading participants to overlook contextual factors like stock identities. Another explanation could relate simply to the inherent difficulty of converting contextual information into financial prices [27]. A final explanation could stem from the design of this study. In particular, all the participants knew they were forecasting stock prices; subjects in the no-labels group did not know which particular stocks were being forecast, while the other participants knew the stock names. Subjects indicated that, when no specific contextual information was given, they did not attempt to identify the particular stocks, but rather tried to base their forecasts on the price movements they could detect as well as their general expectations about the stock market. Given that this experiment was conducted in a highly volatile setting (i.e. prior to national elections),[8] it could be that the wide swings in prices preempted any effects that knowledge of stock names could potentially have on assessors' reactions to feedback. In fact, our analyses clearly reveal the prevailing effects of forecasting session on predictive performance. Taken together, these findings attest to the importance of market volatility on the quality of judgmental predictions, regardless of the elicitation format utilized. Future research investigating the influence of environmental factors like volatility is definitely needed to enhance our understanding of judgmental forecasting.

Post-experimental interviews indicated that all participants found the task very appealing, and yet highly difficult. Overall, subjects who were given calibration feedback expected better probability forecasting performance when compared to subjects receiving outcome feedback. Provision of performance feedback appeared to intensify the focus on performance, leading assessors to closely track their accomplishments across sessions, raising their performance expectations. It is also worth noting that the participants not given label information found it more difficult to make probability forecasts. Although no differences in difficulty were expressed for the interval forecasts, assessment of probabilities were perceived to be easier when stock names were supplied. These accounts suggest "feedback inquiry" [4,5,18,42] as a promising extension of current research. That is, if the participants are to decide on the timing and the type of feedback they would like to access (if any), would there be any resulting differences on forecasting accuracy; and how would the availability of contextual information affect all these considerations?

[8] Out of the 30 stocks being forecasted, 21 stocks (70%) increased in price in session 1, while 4 stocks (13.3%) showed a price increase in session 2, followed by 11 stocks (36.7%) increasing in price in session 3.

Further studies investigating the effects of differing contextual cues [38] and other types of feedback like task properties feedback [37] can be expected to enhance our understanding of the processes involved in judgmental forecasting. Such work may particularly benefit from employing participants with varying levels of expertise [46] and studying combined or group judgments [36,43]. Future research exploring forecasters' use of information and feedback will also be instrumental in designing effective forecast support systems that address users' concerns [12,49]. Financial settings provide ideal platforms for pursuing these issues, with their intrinsically complex, information-rich and dynamic contexts. This complexity, coupled with forecasters' boundless needs for refined predictive accuracy, means that financial forecasting remains an interesting and potent challenge for decision support systems research.

Appendix A. Performance measures for judgmental forecasts

A.1. Hit rate

The hit rate HR_{im} attained by the interval forecasts of subject i for session m is the percentage of the 30 intervals given by the subject (one interval for each stock) that encompasses the realized closing price for that session. From a calibration perspective, it is desirable for the hit rate to be close to the specified 90% confidence coefficient for the relevant intervals. A hit rate below 90% indicates the subject exhibits overconfidence in setting the prediction intervals and a hit rate above 90% indicates the subject exhibits underconfidence.

A.2. Root mean probability score

The probability score PS_{sim} for stock s, subject i, at session m, is defined as the square of the difference between the probability response and the outcome index. That is:

$$PS_{sim} = (f_{sim} - d_{sim})^2$$

The mean MPS_{im} of the probability scores computed for all 30 stocks gives a measure of a subject's overall probability forecasting accuracy, with lower scores suggesting better overall performance. The measure is defined as:

$$MPS_{im} = \frac{1}{30}\sum_{s=1}^{30} PS_{sim}$$

In this study, the square root of MPS_{im} is used as this gives a measure in the form of the original probability units rather than the square of the units; thus $RMPS_{im} = \sqrt{MPS_{im}}$. The best possible score for the MPS and RMPS is zero with an RMPS value of 0.5 providing a benchmark indicative of the performance of the random walk forecaster.

The MPS is associated with Brier [11] and is often referred to as the 'Brier Score'. The MPS is a measure of overall accuracy that can be decomposed in order to highlight unique aspects of judgment yielding critical information about various aspects of forecasting accuracy.

Appendix B. ANOVA model and testing for normality

B.1. ANOVA model

Specifically, the model assumed was

$$Y_{ijkm} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \gamma_{i(jk)} + \delta_m + (\alpha\delta)_{jm} + (\beta\delta)_{km} + (\alpha\beta\delta)_{jkm} + \varepsilon_{(ijkm)}$$

where Y_{ijkm} = mean score across the 30 stocks of the ith subject, feedback level j, label level k, session m; α_j: *feedback* effect, $j=1$ for *outcome* feedback, $j=2$ for *calibration* feedback; β_k: *label* effect, $k=1$ for *no labels* provided, $k=2$ for *labels* provided; $(\alpha\beta)_{jk}$: *feedback* × *label* interaction; $\gamma_{i(jk)}$: *subject* effect (subjects nested within levels of feedback and labels), $i=1,2,3,\ldots,n_{jk}$; δ_m: *session* effect, $m=1, 2, 3$ for forecasting sessions 1, 2, 3; $(\alpha\delta)_{jm}$: *feedback* × *session* interaction; $(\beta\delta)_{km}$: *labels* × *session* interaction; $(\alpha\beta\delta)_{jkm}$: *feedback* × *labels* × *session* interaction; μ = constant (overall mean), α_j's are constants such that $\Sigma\alpha_j = 0$, β_k's are constants such that $\Sigma\beta_k = 0$, δ_m's are constants such that $\Sigma\delta_m = 0$, $(\alpha\beta)_{jk}$'s are constants such that $\Sigma\Sigma(\alpha\beta)_{jk} = 0$, $\gamma_{i(jk)}$'s are constants such that $\Sigma\Sigma\Sigma\gamma_{i(jk)} = 0$, $(\alpha\delta)_{jm}$'s are constants such that $\Sigma\Sigma(\alpha\delta)_{jm} = 0$, $(\beta\delta)_{km}$'s are constants such that $\Sigma\Sigma(\beta\delta)_{km} = 0$, and $(\alpha\beta\delta)_{jkm}$'s

are constants such that $\Sigma\Sigma\Sigma(\alpha\beta\delta)_{jkm} = 0$, $\varepsilon_{(ijkm)} \sim N(0,\sigma^2)$.

B.2. Testing the normality assumption

To examine the normality assumption of the error terms in the model, the normality test [23] was applied to the residuals. Roots of the mean probability score were taken due to apparent non-normalities that were observed resulting in the adoption of RMPS instead. Test results were then consistent with the null hypothesis of normality.

References

[1] L. Adelman, The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks, Organizational Behavior and Human Performance 27 (1981) 423–442.

[2] M. Alpert, H. Raiffa, A progress report on the training of probability assessors, in: D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, Cambridge Univ. Press, Cambridge, 1982, pp. 294–305.

[3] H.R. Arkes, K.R. Hammond (Eds.), Judgment and Decision Making: An Interdisciplinary Reader, Cambridge Univ. Press, Cambridge, 1986.

[4] S.J. Ashford, L.L. Cummings, Feedback as an individual resource: personal strategies of creating information, Organizational Behavior and Human Performance 32 (1983) 370–398.

[5] S.J. Ashford, A.S. Tsui, Self-regulation for managerial effectiveness: the role of active feedback seeking, Academy of Management Journal 34 (1991) 251–280.

[6] W.K. Balzer, M.E. Doherty, R. O'Connor, Effects of cognitive feedback on performance, Psychological Bulletin 106 (1989) 410–433.

[7] J.A. Bartos, The assessment of probability distributions for future security prices, Unpublished PhD thesis, Indiana University, Graduate School of Business, 1969.

[8] R. Batchelor, P. Dua, Forecaster ideology, forecasting technique, and the accuracy of economic forecasts, International Journal of Forecasting 6 (1990) 3–10.

[9] L.R. Beach, V.E. Barnes, J.J.J. Christensen-Szalanski, Beyond heuristics and biases: a contingency model of judgmental forecasting, Journal of Forecasting 5 (1986) 143–157.

[10] P.G. Benson, D. Önkal, The effects of feedback and training on the performance of probability forecasters, International Journal of Forecasting 8 (1992) 559–573.

[11] G.W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Review 78 (1950) 1–3.

[12] D.W. Bunn, Synthesis of expert judgment and statistical forecasting models for decision support, in: G. Wright, F. Bolger (Eds.), Expertise and Decision Support, Plenum, New York, 1992, pp. 251–268.

[13] H. Einhorn, R.M. Hogarth, Prediction, diagnosis, and causal thinking in forecasting, Journal of Forecasting 1 (1982) 23–36.

[14] K.R. Hammond, Computer graphics as an aid to learning, Science 172 (1971) 903–908.

[15] M. Henrion, G.W. Fischer, T. Mullin, Divide and conquer? Effects of decomposition on the accuracy and calibration of subjective probability distributions, Organizational Behavior and Human Decision Processes 55 (1993) 207–227.

[16] P.G.W. Keen, M.S. Scott Morton, Decision Support Systems: An Organizational Perspective, Addison-Wesley, Reading, 1978.

[17] G. Keren, Calibration and probability judgments: conceptual and methodological issues, Acta Psychologica 77 (1991) 217–273.

[18] J.R. Larson, The dynamic interplay between employees' feedback-seeking strategies and supervisors' delivery of performance feedback, Academy of Management Review 14 (1989) 408–422.

[19] M. Lawrence, S. Makridakis, Factors affecting judgmental forecasts and confidence intervals, Organizational Behavior and Human Decision Processes 42 (1989) 172–187.

[20] M. Lawrence, M. O'Connor, Scale, variability and the calibration of judgmental prediction intervals, Organizational Behavior and Human Decision Processes 56 (1993) 441–458.

[21] S. Lichtenstein, B. Fischhoff, Do those who know more also know more about how much they know? Organizational Behavior and Human Performance 20 (1977) 159–183.

[22] S. Lichtenstein, B. Fischhoff, L. Phillips, Calibration of probabilities: the state of the art to 1980, in: D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, Cambridge Univ. Press, Cambridge, 1982, pp. 306–334.

[23] H.W. Lilliefors, On the Kolmogorov–Smirnov test for normality with mean and variance unknown, Journal of the American Statistical Association 62 (1967) 339–402.

[24] J.S. Lim, M. O'Connor, Judgemental adjustment of initial forecasts: its effectiveness and biases, Journal of Behavioral Decision Making 8 (1995) 149–168.

[25] A.G.R. McClelland, F. Bolger, The calibration of subjective probabilities: theories and models 1980–94, in: G. Wright, P. Ayton (Eds.), Subjective Probability, Wiley, Chichester, 1994, pp. 453–482.

[26] G. Muradoglu, D. Önkal, An exploratory analysis of the portfolio managers' probabilistic forecasts of stock prices, Journal of Forecasting 13 (1994) 565–578.

[27] J.J. Murphy, Technical Analysis of Futures Markets, Prentice-Hall, New York, 1986.

[28] J.J. Murphy, Technical Analysis of Financial Markets, New York Institute of Finance, Paramus, NJ, 1999.

[29] M. O'Connor, M. Lawrence, An examination of the accuracy of judgmental confidence intervals in time series forecasting, Journal of Forecasting 8 (1989) 141–155.

[30] M. O'Connor, M. Lawrence, Time series characteristics and the widths of judgmental confidence intervals, International Journal of Forecasting 7 (1992) 413–420.

[31] M. O'Connor, W. Remus, K. Griggs, Judgemental forecasting in times of change, International Journal of Forecasting 9 (1993) 163–172.
[32] D. Önkal, G. Muradoglu, Evaluating probabilistic forecasts of stock prices in a developing stock market, European Journal of Operational Research 74 (1994) 350–358.
[33] D. Önkal, G. Muradoglu, Effects of feedback on probabilistic forecasts of stock prices, International Journal of Forecasting 11 (1995) 307–319.
[34] D. Önkal, G. Muradoglu, Effects of task format on probabilistic forecasting of stock prices, International Journal of Forecasting 12 (1996) 9–24.
[35] D. Önkal-Atay, Financial forecasting with judgment, in: G. Wright, P. Goodwin (Eds.), Forecasting with Judgment, Wiley, Chichester, 1998, pp. 139–167.
[36] S. Plous, A comparison of strategies for reducing interval overconfidence in group judgment, Journal of Applied Psychology 80 (1995) 443–454.
[37] W. Remus, M. O'Connor, K. Griggs, Does feedback improve the accuracy of recurrent judgmental forecasts? Organizational Behavior and Human Decision Processes 66 (1996) 22–30.
[38] D.M. Sanbonmatsu, F.R. Kardes, S.S. Posavac, D.C. Houghton, Contextual influences on judgment based on limited information, Organizational Behavior and Human Decision Processes 69 (1997) 251–264.
[39] J.A. Sniezek, The role of labels in cue probability learning tasks, Organizational Behavior and Human Decision Processes 38 (1989) 141–161.
[40] J.A. Sniezek, T. Buckley, Confidence depends on level of aggregation, Journal of Behavioral Decision Making 4 (1991) 263–272.
[41] C.-A.S. Staël von Holstein, Probabilistic forecasting: an experiment related to the stock market, Organizational Behavior and Human Performance 8 (1972) 139–158.
[42] J.B. Vancouver, E.W. Morrison, Feedback inquiry: the effect of source attributes and individual differences, Organizational Behavior and Human Decision Processes 62 (1995) 276–285.
[43] T.S. Wallsten, D.V. Budescu, I. Erev, A. Diederich, Evaluating and combining subjective probability estimates, Journal of Behavioral Decision Making 10 (1997) 243–268.
[44] M.E. Wilkie, A.C. Pollock, An application of probability judgement accuracy measures to currency forecasting, International Journal of Forecasting 12 (1996) 25–40.
[45] M.E. Wilkie-Thomson, D. Önkal-Atay, A.C. Pollock, Currency forecasting: an investigation of extrapolative judgment, International Journal of Forecasting 13 (1997) 509–526.
[46] G. Wright, G. Rowe, F. Bolger, J. Gammack, Coherence, calibration and expertise in judgmental probability forecasting, Organizational Behavior and Human Decision Processes 57 (1994) 1–25.
[47] I. Yaniv, D. Foster, Precision and accuracy of judgmental estimation, Journal of Behavioral Decision Making 10 (1997) 21–32.
[48] J.F. Yates, L.S. McDaniel, E.S. Brown, Probabilistic forecasts of stock prices and earnings: the hazards of nascent expertise, Organizational Behavior and Human Decision Processes 49 (1991) 60–79.
[49] J.F. Yates, P.C. Price, J.-W. Lee, J. Ramirez, Good probabilistic forecasters: the 'consumer's' perspective, International Journal of Forecasting 12 (1996) 41–56.

Paul Goodwin is Senior Lecturer in Management Science at the University of Bath, UK. His research interests focus on the role of judgment in forecasting and decision making, and he received his PhD from Lancaster University in 1998. He is a co-author of *Decision Analysis for Management Judgment* and a co-editor of *Forecasting with Judgment*, both published by Wiley. He has published articles in a number of academic journals including the *International Journal of Forecasting*, the *Journal of Forecasting*, the *Journal of Behavioral Decision Making*, the *Journal of Multi-Criteria Decision Analysis* and *Omega*.

Dilek Önkal-Atay is an Associate Professor of Decision Sciences and Associate Dean of the Faculty of Business Administration at Bilkent University, Turkey. Dr. Önkal-Atay received her PhD in Decision Sciences from the University of Minnesota. Her research interests include judgmental forecasting, probabilistic financial forecasting, judgment and decision making, decision support systems, risk perception and risk communication. Her work has appeared in several book chapters and journals such as the *International Journal of Forecasting*, the *Journal of Behavioral Decision Making*, the *Journal of Forecasting*, the *International Federation of Technical Analysts Journal*, the *International Forum on Information and Documentation* and the *European Journal of Operational Research*.

Mary E. Thomson is a Reader in the Division of Risk, Glasgow Caledonian Business School. She completed a PhD on judgement in currency forecasting and has published articles and papers in a variety of books and journals in this area. Her research interests focus on the role of judgement in financial forecasting and decision making.

Andrew C. Pollock is a Reader in the School of Computing and Mathematical Sciences, Glasgow Caledonian University. He completed a PhD on exchange rates and has published widely in this area. His particular research interest is the application of analytical techniques to the forecasting of exchange rates and, more generally, financial time series.

Alex Macaulay is Senior Lecturer in Statistics in the School of Computing and Mathematical Sciences, Glasgow Caledonian University. He has a BSc in Mathematics and an MSc in Statistics (Stochastic Processes). He has a particular research interest in the forecasting of financial price series and in the evaluation of predictive performance.